WEKA-based machine learning for traffic congestion prediction in Amman City

Areen Arabiat[1], Mohammad Hassan[1], Omar Almomani[2]
[1]Department of Communication and Computer Engineering, Faculty of Engineering, Al Ahliyya Amman University, Amman, Jordan
[2]Department of Networks and Cybersecurity, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan

Article Info

Article history:

Received Jan 3, 2024
Revised Feb 1, 2024
Accepted Mar 14, 2024

Keywords:

Algorithm
K-nearest neighbor
Logistic regression
Machine learning
Multi-layer perceptron
Support vector machine

ABSTRACT

Traffic congestion leads to wasted time, pollution, and increased fuel consumption. Traffic congestion prediction has become a developing research topic in recent years, particularly in the field of machine learning (ML). The evaluation of various traffic parameters is used to predict traffic congestion by relying on historical data. In this study, we will predict traffic congestion in Amman City, specifically at the 8th circle, using different ML classifiers. The 8th circle links four main streets: Westbound, Northbound, Eastbound, and Southbound. Datasets were collected from the greater Amman municipality hourly. The logistic regression (LR), k-nearest neighbor (KNN), decision tree (DT), random forest (RF), support vector machine (SVM), and multi-layer perceptron (MLP) classifiers have been chosen to predict traffic congestion at each street linked with the 8th circle. The waikato environment for knowledge analysis (WEKA) data mining tool is used to evaluate chosen classifiers by determining accuracy, F-measure, sensitivity, and precision evaluation metrics. The results obtained from all experiments have demonstrated that SVM is the best classifier to predict traffic congestion. The accuracy of SVM to predict traffic congestion at Westbound Street, Northbound Street, Eastbound Street, and Southbound Street was 99.4%, 99.7%, 99.6%, and 99.1%, respectively.

Corresponding Author:

Areen Arabiat
Department of Communications and Computer Engineering, Faculty of Engineering
Al-Ahliyya Amman University
Al-Saro, Al-Salt, Amman, Jordan
Email: a.arabiat@ammanu.edu.jo

1. INTRODUCTION

As cities grow and people move from rural to urban areas, traffic congestion may increase, leading to decreased productivity, air pollution, and increased fuel consumption [1]. Congestion is a major transportation concern in industrial countries, costing $121 billion in the US alone. Inefficient transportation, greenhouse gas emissions, and a lower quality of life for city dwellers are all consequences of traffic congestion [2].

Congestion has been characterized in a variety of ways by researchers. The most common concept of traffic congestion is when the demand for transport exceeds the capacity of the roads. Despite significant improvements in transportation infrastructure, traffic congestion continues to be a major societal and policy problem [3]. In general, traffic congestion can be classified into two types: recurrent congestions, which are usually caused by a mobility demand, and intermittent congestions, which are usually caused by a lack of mobility that surpasses the road network's capacity, including nonrecurring traffic jams caused by incidents or special events [4]. To reduce congestion, three options are suggested: enhancing infrastructure, promoting public transportation in large cities, and predicting future road segments' status. However, these options may require significant investment and may not always be feasible [5]. Air pollution and sustainability are both badly impacted by traffic congestion on road networks. Efficient traffic control can aid in lowering pollution levels. The proliferation of internet of things (IoT) devices provides data sets for intelligent, environmentally friendly transportation options. Depending on the design of the roads, long short-term memory (LSTM) networks can forecast the spread of congestion over a network of roads with an accuracy of 84–95%. This may be a crucial part of traffic modeling in the future for smart cities [6]. Traffic congestion prediction has grown significantly in recent decades due to big data from sensors and new artificial intelligence (AI) models. Machine learning (ML), a part of AI, is used to evaluate traffic parameters and predict short-term congestion. AI has made significant advances in ML, data mining, computer vision, expert systems, natural language processing, and robotics. Prediction issues can be classified as ML classifications or regression models [7].

Moumen *et al.* [8] shows how combining AI and IoT road traffic data might enhance urban mobility through traffic prediction in smart city settings. The LSTM model enhances real-time vehicle count estimates and projects future traffic patterns using IoT sensors and deep learning algorithms, allowing for well-informed decision-making. The potential benefits of integrating IoT and AI are enormous, including improved urban mobility, less traffic, less congestion, and effective traffic management. In this study, LSTM accuracy achieved 91%, whereas linear regression, k-nearest neighbor (KNN), and support vector machine (SVM) accuracy reached 41%, 43%, and 46% of the total, respectively. Najm *et al.* [9] identified several ML algorithms in order to predict the best congestion control for 5G IoT wireless sensors. This study suggests a novel ML model based on the decision tree (DT) method. Over 92% accuracy and recall were attained by the model after it was trained on a training dataset. In Helsinki, Finland, a study introduced the critical path method using convolutional long short-term memory (CPM-ConvLSTM), a spatiotemporal model that predicts congestion levels in each road segment in the short run outperforms six rivals in prediction accuracy based on traffic data [10]. Using a driver model that uses SVM to anticipate traffic congestion without relying on traffic flow monitoring technology. The model uses steering, throttle, and speed input frequency changes in driving simulators, requiring no additional sensors or infrastructure [11].

A traffic congestion prediction model is developed using random forest (RF), a robust and high-performance ML approach. The model, incorporating weather, period, special road conditions, road quality, and holidays, achieved an accuracy of 87.5% and a low generalization error [12]. Congestion matrices using different methods are developed on regional traffic networks using relative positions for road nodes. They used a convolutional long-short-term memory network to predict congestion in all sections of the network. The approach outperformed baseline models and accurately captured traffic's temporal and spatial characteristics, demonstrating its interpretability for congestion prediction [13]. Another study used seven ML algorithms to predict traffic congestion using a binary classification problem. The algorithms included KNN, DT, artificial neural network (ANN), stochastic gradient descent (SGD), fuzzy unordered rule induction algorithm (FURIA), Bayesian network (BN), and SVM. Ensemble learning algorithms like bagging, boosting, piling, and PTS were used to enhance the prediction accuracy [14]. Using the logistic regression (LR) technique in [15] reveals that it has a 91% accuracy rate in traffic analysis, providing the quickest and most direct route to desired locations. This method reduces travel time, noise pollution, carbon dioxide emissions, and fuel consumption. Linear LR predicts probability values using a linear combination of features, with chances ranging from zero to infinity [15]. The predicting of traffic congestion using regression models are becoming increasingly ineffective as the amount of data and its complexity rise. Mapping nonlinear data to a high-dimensional linear space where it may be linearly categorized using hyperplanes is the fundamental concept of SVM [16], [17].

Modern ML algorithms and data preprocessing tools are arranged in an orderly manner on the Weka workbench. Using these methods from the command line is the primary method of interacting with them. But there are also easy-to-use interactive graphical user interfaces available for data exploration, large-scale experiment setup on distributed computing platforms, and streamed data processing configuration design. These interfaces provide a sophisticated setting for data mining experiments. The GNU general public license governs the distribution of the Java-written system [18].

This study aims to predict congestion using AI techniques, especially machine and deep learning, using the Weka tool for data that was collected from Greater Amman Municipality. The data were cleaned, processed, and inserted into the WEKA tool using the selected classifiers. Using a variety of statistical metrics, including accuracy, specificity, sensitivity, and the F1-measure, classifiers were employed to find the best one to predict traffic congestion on all roads that enter the eighth circle in Amman City. These matrices make it possible to identify congestion more precisely. Then the best classifier will be determined according to the highest accuracy, reliability, perception, and F-massure to predict traffic congestion precisely.

In order to improve the prediction, a novel approach involves utilizing a huge amount of traffic data from four approaches that entering study area. Additionally, by using six classifiers including LR, KNN, DT, RF, SVM, and multi-layer perceptron (MLP) an extensive review and twenty-four experiments are carried out to identify the best classifier with the highest performance. On the other hand, the data was collected over a one-year period, and an ML model was constructed using WEKA data minig which is different from previous research.

2. METHOD

2.1. Traffic congestion prediction system architecture

The proposed model aims to predict traffic congestion at Amman's 8th circle intersection using different ML algorithms. Weka will be utilized to develop ML models in this methodology section. At the beginning, the dataset was obtained from Greater Amman Municipality, then it was preprocessed and converted to a comma-separated value (CSV) to be readable by WEKA.Then, the dataset is fed to WEKA to build the required model using SVM, MLP, LR, DT, RF, and KNN classifiers. Random sampling with a 70% training and a 30% testing set size was able to get a good result because there were a lot of records (8640) in the dataset for each bound. Also, 10-fold cross-validation is applied to increase the performance. Using the confusion matrix that results from WEKA, the dataset is evaluated in terms of accuracy, f-measure, precision, and recall after the trained model has been constructed. Furthermore, a comparative analysis will be conducted to ascertain which ML model exhibits the highest performance indicators. Findings show which ML algorithms are the best to predict traffic congestion, among others. Figure 1 depicts the traffic congestion prediction system architecture.

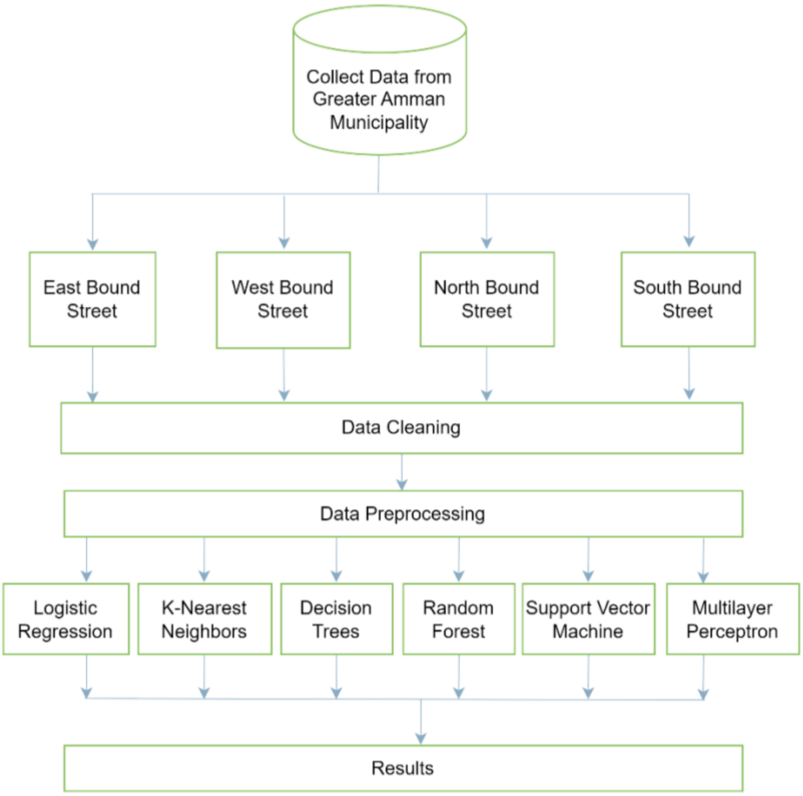
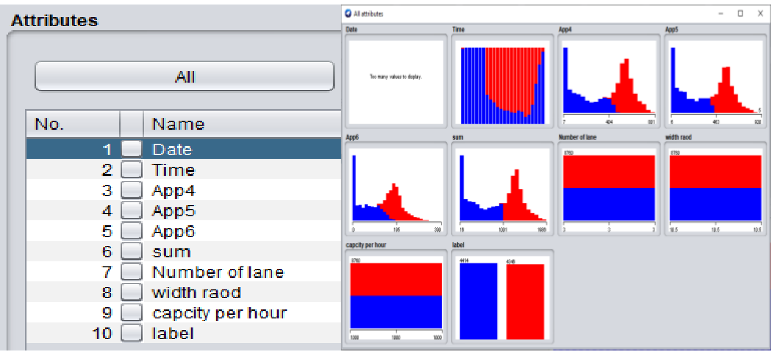


Figure 1. Traffic congestion prediction system architecture

2.2. Dataset

The dataset, obtained from the Greater Amman Municipality for the period from 1/1/2019 to 31/12/2019, was chosen for efficient and precise traffic congestion prediction. It includes traffic volume for each lane, density, speed, occupancy, width, and distance. Traffic volume is determined for each approach on each lane using detectors and sensors. Traffic volume was listed for each approach for 24 hours, every month, for the year 2019. The list of attributes for congestion prediction analysis and attribute list visualization are shown in Figure 2. However, the dataset includes the data of all approaches entering the 8th circle, as shown in Figure 3, including westbound, northbound, eastbound, and southbound.



Attributes

All

No.	Name
1	□ Date
2	□ Time
3	□ App4
4	□ App5
5	□ App6
6	□ sum
7	□ Number of lane
8	□ width raod
9	□ capcity per hour
10	□ label

Figure 2. WEKA's list of attributes for congestion prediction analysis and attributes list visualization

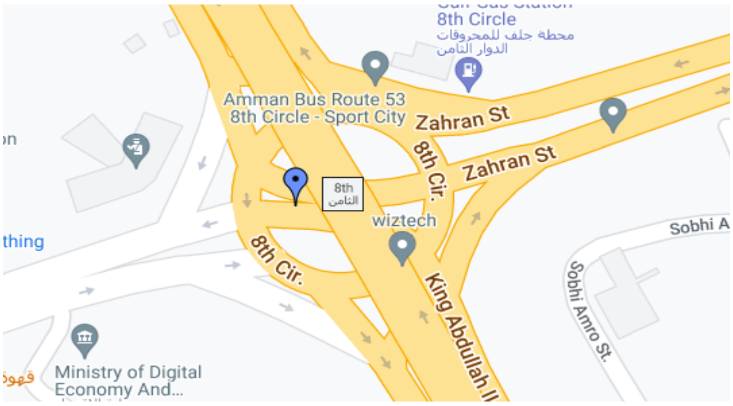


Figure 3. Study area [19]

The next step is data cleansing, which involves correcting or deleting incorrect, corrupted, improperly formatted, duplicate, or incomplete data from a dataset. In our traffic congestion prediction system, redundant data was removed using the remove duplicates feature in Excel. Structural errors were fixed by adding a new rule using conditional formatting to correct incorrect classifications. The third step is data preprocessing, which is the crucial first step in the development of an ML classifier, ensuring the data meets the analysis needs. The preparation module in WEKA handles this process, typically saving the traffic dataset as a CSV file. Figure 3 depicts a list of attributes and classes for congestion prediction analysis; Figure 4 shows the attribute list visualization of WEKA.

2.3. Machine learning classifiers

ML is a technique for teaching machines to handle data better. ML is increasingly popular due to the abundance of datasets and is used in various businesses to extract important data [20]. ML has been crucial in smart transportation, investigating complex interactions between roadways, traffic, environmental elements, and traffic crashes. ML is used in applications like data mining, image processing, and predictive analytics as algorithms learn to complete tasks independently [21].

2.3.1. Logistic regression

LR is a fundamental classification technique, a linear classifier that uses probability to categorize data into binary groups [22]. It is a simple and quick method for data analysis, allowing for easy understanding of findings. It primarily focuses on binary classification but can also solve multiclass issues [23]. The basic model of the LR estimation is as (1):

$$\frac{Prob(Yi=1)}{Prob(Yi=0)} = \frac{Pi}{1-Pi} = e(\ β0+\ β1\ X1+\ldots\ldots+\ βk\ Xki) \qquad (1)$$

where e is the exponential constant, (1-Pi) is the chance that Y takes a value of 0, and Pi is the probability that Y takes a value of 1 [24].

2.3.2. K-nearest neighbor

A KNN classifier is a reliable data classification algorithm, but its accuracy depends on the choice of nearest KNNs [25]. It ignores the k-environment distribution, making constant k numbers unsuitable for irregular datasets. Dynamic k selection is suggested, but maximizing performance is challenging for large datasets [26]. KNN excels in classification due to its simplicity. However, large sample sizes and feature attributes can hinder its performance. The parameter k, which determines the number of KNNs, significantly impacts the algorithm's diagnostic performance, requiring a balance between overfitting and underfitting [27].

2.3.3. Decision tree

DT classifiers are widely used data classification methods in various domains, including ML, image processing, pattern recognition, and traffic congestion prediction. They consist of nodes and branches and use various classification algorithms to manage missing values' continuous and periodic properties [28]. Interpretable models (DT) are widely used for categorization and decision-making, but their myopic induction algorithms lead to poor predictive performance and fundamental biases when intricate interactions between input features occur [28], [29].

2.3.4. Random forest

RF classifier is a random method that creates multiple DTs using a random vector to eliminate correlations and improve accuracy. Each DT is divided into a subset of characteristics, with the number of traits considered affecting the tree variety. The ideal global function is found at each break, allowing for similar trees. The goal is to create a mixture of decision-making trees for different forecasts [30]–[32].

2.3.5. Support vector machine

SVMs are supervised learning algorithms that may be applied to both classification and regression applications. The basic goal of SVMs is to discover the hyperplane that best separates the classes in the feature space while decreasing the amount of misclassifications. This is accomplished by solving an optimization problem that aims to identify the best hyperplane that balances the margin (the distance between the hyperplane and the nearest data points) and the error rate. Therefore, SVMs excel at handling high-dimensional data and are frequently favored over other ML algorithms when working with non-linearly separable datasets [33], [34].

2.3.6. Multi-layer perceptron

MLP is a deep learning neural network with a three-layer structure, including the input layer, hidden layer or layers, and output layer or layers, in which every neuron is coupled to every other neuron in the layer above, which is beneficial for MLP, as Figure 4 illustrates. MLP using back-propagation and error gradient propagation techniques for data transfer and error gradient propagation for training highlights that MLP produces high-quality models with short training times but requires a modular model for multiple output values. Independent training evaluates all architectures, increasing the chances of finding a superior prediction model [35]–[37].

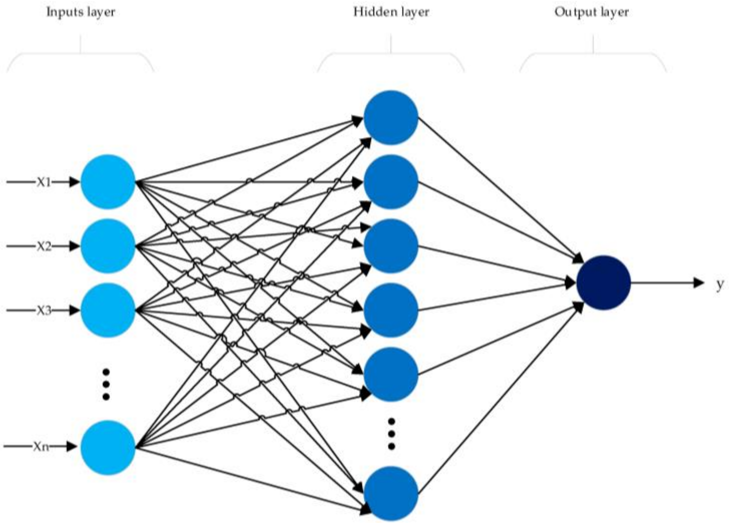


Figure 4. Multilayer perceptron neural networks architecture [38]

2.4. WEKA tool

ML and data mining can be facilitated by using the open-source WEKA toolbox. The University of Waikato in New Zealand developed WEKA, a user-friendly graphical user interface offering a full array of training, assessment, and data preparation tools as shown in Figure 5. Researchers and practitioners may extract useful insights and information from their data using WEKA's ability to handle a wide range of data types and sources, such as databases, spreadsheets, and text files. WEKA also supports a number of ML methods, such as SVM, DT, and neural networks, giving users the option to choose the one that best suits their requirements. For examining and evaluating large, complicated data sets, WEKA is an all-around strong and adaptable platform [39]. There are a number of features available in WEKA's graphical user interface (GUI), such as the Explorer that lets researchers access different facilities, the experimenter that compares the predictive performance of learning algorithms on a large scale, the knowledge flow interface that lets users layout components like filters, classifiers, and evaluations interactively, the workbench that combines all other WEKA GUIs, and the simple CLI that lets users execute WEKA commands directly [40], [41]. Figure 5 depicts the WEKA tools GUI.

Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand

Figure 5.WEKA tools GUI

3. CLASSIFICATION AND RESULT IMPLEMNTAION

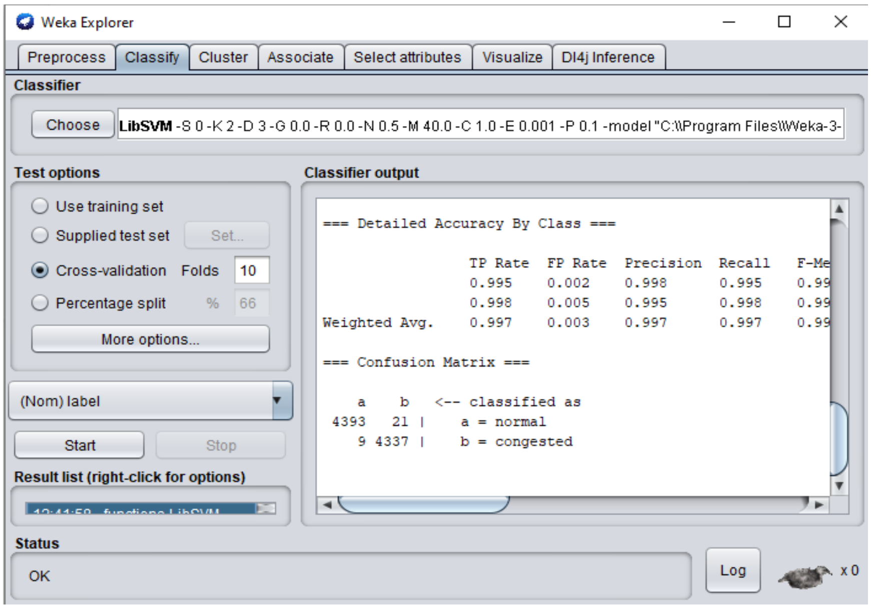
The 8th Circle connects four main streets in Amman, with each street running as a separate experiment. The first experiment consists of three approaches from Westbound Street (1, 2, and 3), collected from the greater Amman Municipality hourly from 1-1-2019 to 31-12-2019. The second experiment uses three approaches from Northbound Street (4,5,6), collected from the greater Amman Municipality hourly from 1-1-2019 to 31-12-2019. The third experiment uses three approaches from Eastbound Street (7, 8, and 9), collected from the greater Amman Municipality hourly from 1-1-2019 to 31-12-2019. The fourth experiment uses one approach from the airport (10).

3.1. Performance matrices

The following metrics are used to assess the efficiency of the traffic congestion prediction model: true positive(TP), true negative (TN), false positive (FP), and false negative (FN) as shown in Table 1. To evaluate the results of the confusion matrix, accuracy, precision, and recall are methods for summarizing the results [42], [43]. The confusion matrix of SVM using the WEKA interface is shown in Figure 6.

Table 1. Confusion matrix

		Predicted	
		Congested	Uncongested
Actual	Congested	TP	FN
	Uncongested	FP	TN



LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -model "C:\\Program Files\\Weka-3-

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) label

OK

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Me
                 0.995    0.002    0.998      0.995    0.99
                 0.998    0.005    0.995      0.998    0.99
Weighted Avg.    0.997    0.003    0.997      0.997    0.99

=== Confusion Matrix ===

    a    b   <-- classified as
 4393   21 |    a = normal
    9 4337 |    b = congested
```

Figure 6. Result of SVM using WEKA interface

The classification accuracy statement serves as the fundamental criterion for assessing the suitability of a classification system for its intended purpose. Accuracy statements are also employed in various applications, such as classifier evaluation, where particular emphasis is placed on discerning variations in the accuracy of data classification [44]. In other words, it's a measure of how well a model performs across all targets. It’s calculated by dividing the number of right predictions by the total number of predictions, as shown in (2) [45]. Table 2 shows the accuracy comparison of each experiment for different ML classifiers. Figure 7 demonstrates the accuracy of the representation of all classifiers for all experiments.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2)$$

Table 2. Accuracy comparison of all classifiers for all experiments

	DT (%)	SVM (%)	RF (%)	KNN (%)	MLP (%)	LR (%)
Westbound Street	97.7	99.4	97.5	98.0	94.5	96.6
Southbound Street	96.3	99.1	96.9	98.5	95.7	97.1
Eastbound Street	97.2	99.6	97.1	98.3	96.3	96.6
Northbound Street	97.2	99.7	96.2	98.7	95.4	97.6

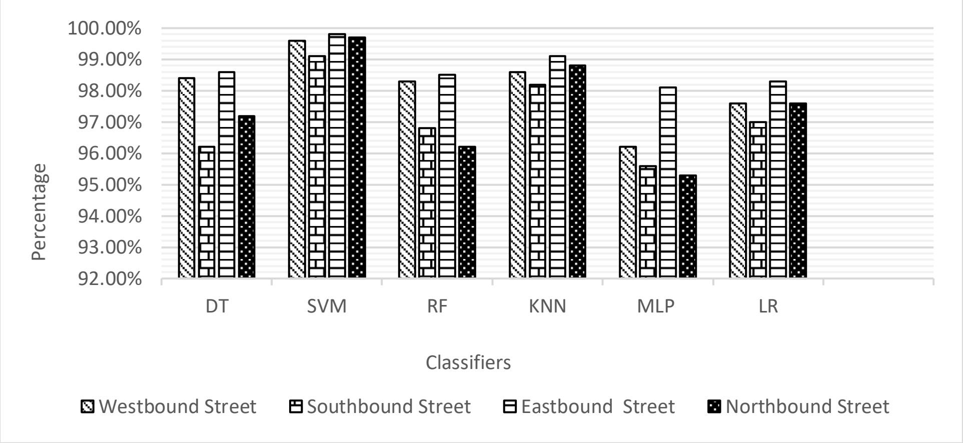


Figure 7. Accuracy comparison of all classifiers for all experiments

In terms of precision, it can be determined by the ratio of correctly identified positive (congested) to total positive (either congested or uncongested), as shown in (3) [46]. Table 3 shows the accuracy comparison of each experiment for different ML classifiers. Figure 8 demonstrates the accuracy representation of all classifiers for all experiments.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{3}$$

Table 3. Precision comparison of all classifiers for all experiments

	DT (%)	SVM (%)	RF (%)	KNN (%)	MLP (%)	LR (%)
Westbound Street	98.4	99.4	98.7	98.6	97.0	99.2
Southbound Street	95.2	98.8	97.2	98.3	96.9	97.8
Eastbound Street	100.0	100.0	100.0	100.0	99.9	100.0
Northbound Street	96.9	99.8	97.0	99.0	98.4	97.7

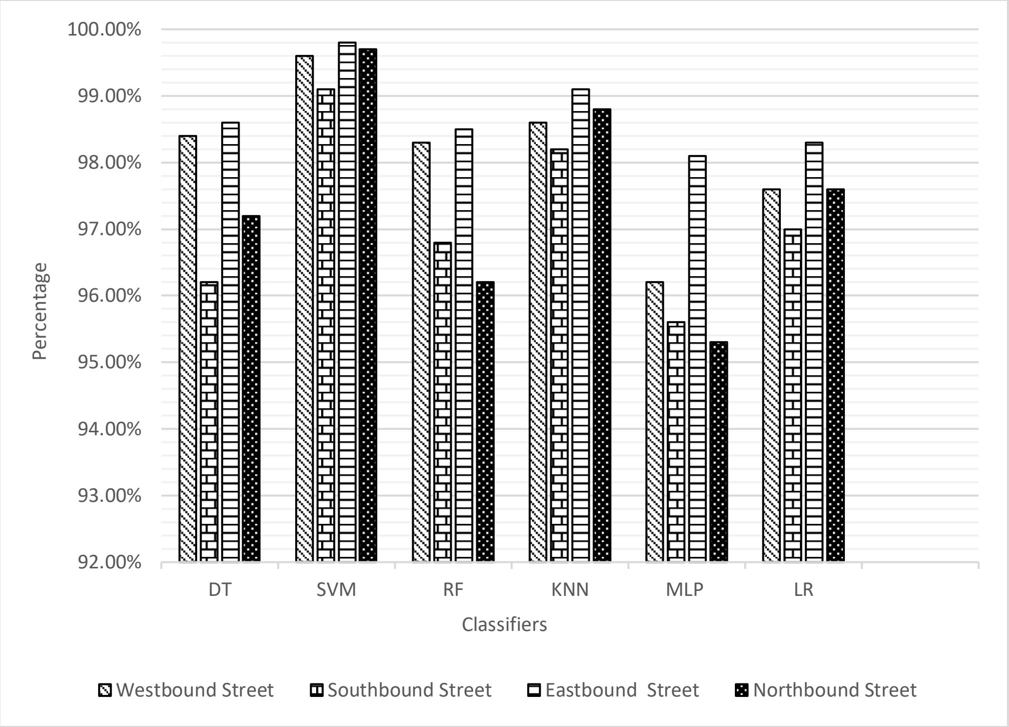


Figure 8. Precision comparison of all classifiers for all experiments

Sensitivity is another performance evaluation used in this study, which is computed as the ratio of correctly identified positive values to the total number of positive values. The sensitivity matrix determines how well the classifier can discover positive values, as shown in (4) [47]. Table 4 shows the accuracy comparison of each experiment for different ML classifiers. Figure 9 demonstrates the accuracy of the representation of all classifiers for all experiments.

$$\text{Sensitivity } = \frac{\text{TP}}{\text{TP+FN}} \tag{4}$$

Table 4. Sensitivity comparison of all classifiers for all experiments

	DT (%)	SVM (%)	RF (%)	KNN (%)	MLP (%)	LR (%)
Westbound Street	98.4	99.8	97.8	98.6	95.3	96.0
Southbound Street	97.2	99.3	96.5	98.1	94.2	96.3
Eastbound Street	97.2	99.6	97.1	98.3	96.4	96.6
Northbound Street	97.6	99.5	95.4	98.5	92.5	97.5

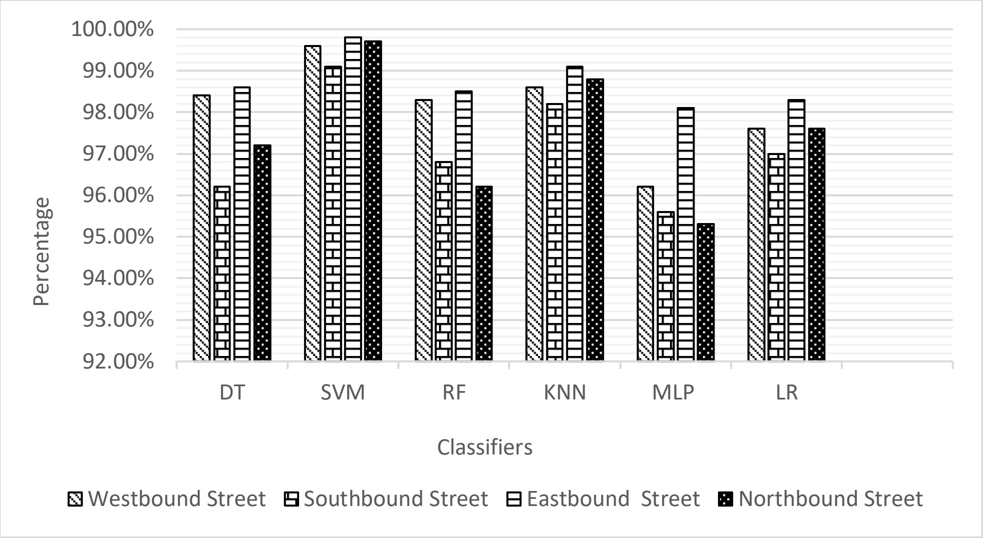


Figure 9. Sensitivity comparison of all classifiers for all experiments

The last performance of the result evaluation that is used in this paper is the F-measure, which is determined by taking the harmonic mean of precision and sensitivity and assigning equal weighting to each as shown in (5) [48]. Table 5 shows the accuracy comparison of each experiment for different ML classifiers. Figure 10 demonstrates the accuracy of the representation of all classifiers for all experiments.

$$\text{F} - \text{Measure} = \frac{2*\text{Precision}*\text{Sensitivity}}{\text{Precision}+\text{Sensitivity}} \quad (5)$$

Table 5. F-measure comparison of all classifiers for all experiments

	DT (%)	SVM (%)	RF (%)	KNN (%)	MLP (%)	LR (%)
Westbound Street	98.4	99.6	98.3	98.6	96.2	97.6
Southbound Street	96.2	99.1	96.8	98.2	95.6	97.0
Eastbound Street	98.6	99.8	98.5	99.1	98.1	98.3
Northbound Street	97.2	99.7	96.2	98.8	95.3	97.6

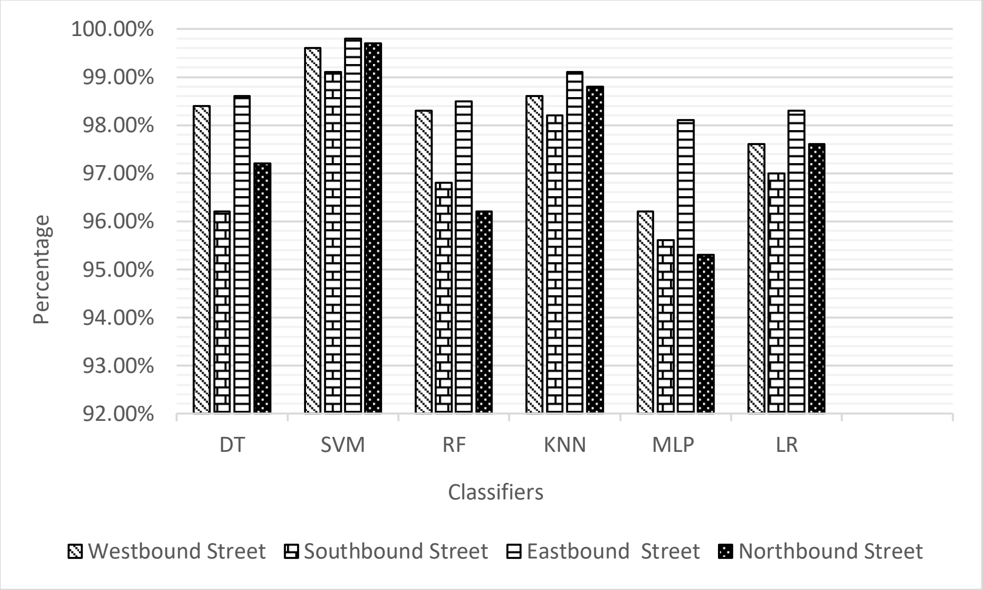


Figure10. F-measure comparison of all classifiers for all experiments

3.2. Results of westbound street

This experiment was conducted using traffic data from Southbound Street. The results were evaluated based on the LR, KNN, DT, RF, SVM, and MLP classifiers. According to the result, the accuracy of the SVM classifier was the highest of all classifiers; it reached 99.8%. In contrast, the lowest accuracy was for MLP, with 94.5%. Table 6 shows the results of west-bound street performance matrices.

Table 6. Westbound street results

	DT (%)	SVM (%)	RF (%)	KNN (%)	MLP (%)	LR (%)
TP	0.984	0.998	0.978	0.986	0.953	0.960
FN	0.016	0.002	0.022	0.014	0.047	0.040
FP	0.040	0.016	0.032	0.035	0.074	0.021
TN	0.960	0.984	0.968	0.965	0.926	0.979
Accuracy	0.977	0.994	0.975	0.980	0.945	0.966
Precision	0.984	0.994	0.987	0.986	0.970	0.992
Sensitivity	0.984	0.998	0.978	0.986	0.953	0.960
F-measure	0.984	0.996	0.983	0.986	0.962	0.976

3.3. Results of northbound street

This experiment was conducted using traffic data from Northbound North. The results were evaluated based on the LR, KNN, DT, RF, SVM, and MLP classifiers. According to the result, the accuracy of the SVM classifier was the highest of all classifiers; it reached 99.7%. In contrast, the lowest accuracy was for MLP, with 95.4%. Table 7 shows the results of the north-bound street performance matrices.

Table 7. Northbound street results

	DT (%)	SVM (%)	RF (%)	KNN (%)	MLP (%)	LR (%)
TP	0.976	0.995	0.954	0.985	0.925	0.975
FN	0.024	0.005	0.046	0.015	0.075	0.025
FP	0.031	0.002	0.030	0.010	0.016	0.023
TN	0.969	0.998	0.970	0.990	0.984	0.977
Accuracy	0.972	0.997	0.962	0.987	0.954	0.976
Precision	0.969	0.998	0.970	0.990	0.984	0.977
Sensitivity	0.976	0.995	0.954	0.985	0.925	0.975
F-measure	0.972	0.997	0.962	0.988	0.953	0.976

3.4. Results of eastbound street

This experiment was conducted using traffic data from Eastbound East. The results were evaluated based on the LR, KNN, DT, RF, SVM, and MLP classifiers. According to the result, the accuracy of the SVM classifier was the highest of all classifiers; it reached 99.6%. In contrast, the lowest accuracy was for MLP, with 96.3%. Table 8 shows the results of east-bound street performance matrices.

Table 8. Eastbound street results

	DT (%)	SVM (%)	RF (%)	KNN (%)	MLP (%)	LR (%)
TP	0.972	0.996	0.971	0.983	0.964	0.966
FN	0.028	0.004	0.029	0.017	0.036	0.034
FP	0.400	0.300	0.300	0.200	0.500	0.300
TN	0.600	0.700	0.700	0.800	0.500	0.700
Accuracy	0.972	0.996	0.971	0.983	0.963	0.966
Precision	1.000	1.000	1.000	1.000	0.999	1.000
Sensitivity	0.972	0.996	0.971	0.983	0.964	0.966
F-measure	0.986	0.998	0.985	0.991	0.981	0.983

3.5. Results of Southbound Street

This experiment was conducted using traffic data from Southbound South. The results were evaluated based on the LR, KNN, DT, RF, SVM, and MLP classifiers. According to the result, the accuracy of the SVM classifier was the highest of all classifiers; it reached 99.1%. In contrast, the lowest accuracy was for MLP, with 95.7%. Table 9 shows the results of the south-bound street performance matrices.

Table 9. Southbound street performance results

	DT (%)	SVM (%)	RF (%)	KNN (%)	MLP (%)	LR (%)
TP	0.972	0.993	0.965	0.981	0.942	0.963
FN	0.028	0.007	0.035	0.019	0.058	0.037
FP	0.046	0.011	0.027	0.012	0.028	0.021
TN	0.954	0.989	0.973	0.988	0.972	0.979
Accuracy	0.963	0.991	0.969	0.985	0.957	0.971
Precision	0.952	0.988	0.972	0.983	0.969	0.978
Sensitivity	0.972	0.993	0.965	0.981	0.942	0.963
F-measure	0.962	0.991	0.968	0.982	0.956	0.970

4. DISSCUSSION

Traffic congestion prediction using ML has great benefits for reducing time waste, fuel consumption, and saving money. In this paper, LR, KNN, DT, RF, SVM, and MLP classifiers were used in the prediction process, and according to the results, SVM had the highest classification accuracy, but MLP had the lowest accuracy. In contrast, SVM has the highest precision score of 98.8%. In terms of sensitivity and F-measure, SVM also has the highest rating score. Finally, by balancing all measures of these classifiers on westbound, northbound, eastbound, and southbound, it can accurately predict both positive and negative classes. SVM had the best accuracy on Northbound (99.8%). On the other hand, the accuracy demonstrated in this paper is shown to be superior to the previous research. Research by Majumdar *et al*. [6], the accuracy rate was 84–95%; in Di *et al*. [10], the accuracy rate was 87.5%; in Moumen *et al*. [8], the maximum accuracy rate was 91%; and in Lakshna *et al*. [15], the accuracy rate was 91%. However, in this study, the best accuracy rate is around 99.7% for Northbound Street.

5. CONCLUSION

This study investigates several classification techniques for predicting traffic congestion in Amman City, the capital of Jordan, specificly the 8th circle area. Four datasets, each one contains approximately 8640 records, were split up and processed for each street. The dataset was gathered from the greater Amman Municipality. In order to predict traffic congestion at each bound connected to the 8th Circle, WEKA data mining is used. The dataset was fed to WEKA to find the confusion matrix in order to determine the best classifier for predicting traffic congestion. LR, KNN, DT, RF, SVM, and MLP classifiers was used in this paper, the accuracy, precision, sensitivity, and F-measure assessment measures have been determine for all selected classifiers to evlautate classifiers prrformance. The findings showed that the SVM is the best classifier for predicting traffic congestion for all bounds. SVM accuracy was 99.4%, 99.7%, 99.6%, and 99.1%, respectively. However, in this research, the best accuracy rate for SVM on Northbound Street was 99.7%.

REFERENCES

[1] M. R. Jabbarpour, H. Zarrabi, R. H. Khokhar, S. Shamshirband, and K. K. R. Choo, "Applications of computational intelligence in vehicle traffic congestion problem: a survey," *Soft Computing*, vol. 22, no. 7, pp. 2299–2320, 2018, doi: 10.1007/s00500-017-2492-z.

[2] H. Xiong, A. Vahedian, X. Zhou, Y. Li, and J. Luo, "Predicting traffic congestion propagation patterns," in *Proceedings of the 11th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, New York, USA: ACM, Nov. 2018, pp. 60–69, doi: 10.1145/3283207.3283213.

[3] T. Afrin and N. Yodo, "A survey of road traffic congestion measures towards a sustainable and resilient transportation system," *Sustainability*, vol. 12, no. 11, Jun. 2020, doi: 10.3390/su12114660.

[4] S. Kwoczek, S. D. Martino, and W. Nejdl, "Predicting and visualizing traffic congestion in the presence of planned special events," *Journal of Visual Languages & Computing*, vol. 25, no. 6, pp. 973–980, Dec. 2014, doi: 10.1016/j.jvlc.2014.10.028.

[5] F. Lécué, R. Tucker, V. Bicer, P. Tommasi, S. T. -Diotallevi, and M. Sbodio, "Predicting severity of road traffic congestion using semantic Web technologies," in *The Semantic Web: Trends and Challenges*, 2014, pp. 611–627, doi: 10.1007/978-3-319-07443-6_41.

[6] S. Majumdar, M. M. Subhani, B. Roullier, A. Anjum, and R. Zhu, "Congestion prediction for smart sustainable cities using IoT and machine learning approaches," *Sustainable Cities and Society*, vol. 64, Jan. 2021, doi: 10.1016/j.scs.2020.102500.

[7] M. Akhtar and S. Moridpour, "A review of traffic congestion prediction using artificial intelligence," *Journal of Advanced Transportation*, vol. 2021, pp. 1–18, Jan. 2021, doi: 10.1155/2021/8878011.

[8] I. Moumen, J. Abouchabaka, and N. Rafalia, "Enhancing urban mobility: integration of IoT road traffic data and artificial intelligence in smart city environment," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 2, pp. 985–993, Nov. 2023, doi: 10.11591/ijeecs.v32.i2.pp985-993.

[9] I. A. Najm, A. K. Hamoud, J. Lloret, and I. Bosch, "Machine learning prediction approach to enhance congestion control in 5G IoT environment," *Electronics*, vol. 8, no. 6, May 2019, doi: 10.3390/electronics8060607.

[10] X. Di, Y. Xiao, C. Zhu, Y. Deng, Q. Zhao, and W. Rao, "Traffic congestion prediction by spatiotemporal propagation patterns," in *2019 20th IEEE International Conference on Mobile Data Management (MDM)*, IEEE, Jun. 2019, pp. 298–303, doi: 10.1109/MDM.2019.00-45.

[11] T. Ito and R. Kaneyasu, "Predicting traffic congestion using driver behavior," *Procedia Computer Science*, vol. 112, pp. 1288–1297, 2017, doi: 10.1016/j.procs.2017.08.090.

[12] Y. Liu and H. Wu, "Prediction of road traffic congestion based on random forest," in *2017 10th International Symposium on*

Computational Intelligence and Design (ISCID), IEEE, Dec. 2017, pp. 361–364, doi: 10.1109/ISCID.2017.216.

[13] M. Bai, Y. Lin, M. Ma, P. Wang, and L. Duan, "PrePCT: Traffic congestion prediction in smart cities with relative position congestion tensor," *Neurocomputing*, vol. 444, pp. 147–157, Jul. 2021, doi: 10.1016/j.neucom.2020.08.075.

[14] G. A. -Cortés, E. Florido, A. Troncoso, and F. M. -Álvarez, "A novel methodology to predict urban traffic congestion with ensemble learning," *Soft Computing*, vol. 20, no. 11, pp. 4205–4216, Nov. 2016, doi: 10.1007/s00500-016-2288-6.

[15] A. Lakshna, K. Ramesh, B. Prabha, D. Sheema, and K. Vijayakumar, "Machine learning smart traffic prediction and congestion reduction," in *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, IEEE, Sep. 2021, pp. 1–4, doi: 10.1109/ICSES52305.2021.9633949.

[16] H. Wang, L. Liu, S. Dong, Z. Qian, and H. Wei, "A novel work zone short-term vehicle-type specific traffic speed prediction model through the hybrid EMD–ARIMA framework," *Transportmetrica B: Transport Dynamics*, vol. 4, no. 3, pp. 159–186, Sep. 2016, doi: 10.1080/21680566.2015.1060582.

[17] Q. Ding, X. Wang, X. Zhang, and Z. Sun, "Forecasting traffic volume with space-time ARIMA model," *Advanced Materials Research*, vol. 156–157, pp. 979–983, 2011, doi: 10.4028/www.scientific.net/AMR.156-157.979.

[18] E. Frank *et al.*, "WEKA-a machine learning workbench for data mining," in *Data Mining and Knowledge Discovery Handbook*, Boston, MA: Springer US, 2009, pp. 1269–1277, doi: 10.1007/978-0-387-09823-4_66.

[19] "8th Circle," *Google Maps*. [Online]. Available: https://www.google.com/maps/d/viewer?mid=1D16ICyhNnv-C8l8rQ_JH2755Icg&hl=en&ll=31.95661328046349%2C35.847913850971246&z=18

[20] R. Li, M. Liu, D. Xu, J. Gao, F. Wu, and L. Zhu, "A Review of machine learning algorithms for text classification," *Communications in Computer and Information Science*, vol. 1506, pp. 226–234, 2022, doi: 10.1007/978-981-16-9229-1_14.

[21] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, Jan. 2021, doi: 10.1016/j.aci.2018.08.003.

[22] M. Stojiljković "Logistic regression in Python," *Real Python*, 2019. Accessed: Jun. 26, 2023. [Online]. Available: https://realpython.com/logistic-regression-python/

[23] M. Hossain *et al.*, "A predictive logistic regression model for periodontal diseases," *Saudi Journal of Oral Sciences*, vol. 8, no. 3, 2021, doi: 10.4103/sjoralsci.sjoralsci_123_20.

[24] E. Y. Boateng and D. A. Abaye, "A review of the logistic regression model with emphasis on medical research," *Journal of Data Analysis and Information Processing*, vol. 7, no. 4, pp. 190–207, 2019, doi: 10.4236/jdaip.2019.74012.

[25] A. Onyezewe, A. F. Kana, F. B. Abdullahi, and A. O. Abdulsalami, "An enhanced adaptive k-nearest neighbor classifier using simulated annealing," *International Journal of Intelligent Systems and Applications*, vol. 13, no. 1, pp. 34–44, Feb. 2021, doi: 10.5815/ijisa.2021.01.03.

[26] Mohebbanaaz, L. V. R. Kumari, and Y. P. Sai, "Classification of arrhythmia beats using optimized k-nearest neighbor classifier," in *Intelligent Systems*, vol. 185, pp. 349–359, 2021, doi: 10.1007/978-981-33-6081-5_31.

[27] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Annals of Translational Medicine*, vol. 4, no. 11, pp. 218–218, Jun. 2016, doi: 10.21037/atm.2016.03.37.

[28] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[29] O. Sagi and L. Rokach, "Approximating XGBoost with an interpretable decision tree," *Information Sciences*, vol. 572, pp. 522–542, 2021, doi: 10.1016/j.ins.2021.05.055.

[30] H. Aydadenta and Adiwijaya, "On the classification techniques in data mining for microarray data classification," *Journal of Physics: Conference Series*, vol. 971, Mar. 2018, doi: 10.1088/1742-6596/971/1/012004.

[31] M. G. Raj, C. Pradip, N. Saju, and S. V. T. Sangeetha, "Random forest-based method for micro grid system in energy consumption prediction," *Journal of Physics: Conference Series*, vol. 1964, no. 5, 2021, doi: 10.1088/1742-6596/1964/5/052002.

[32] P. Palimkar, R. N. Shaw, and A. Ghosh, "Machine learning technique to prognosis diabetes disease: random forest classifier approach," in *Advanced Computing and Intelligent Technologies*, vol. 218, pp. 219–244, 2022, doi: 10.1007/978-981-16-2164-2_19.

[33] A. Tharwat, "Parameter investigation of support vector machine classifier with kernel functions," *Knowledge and Information Systems*, vol. 61, no. 3, pp. 1269–1302, 2019, doi: 10.1007/s10115-019-01335-4.

[34] Y. W. Liu, H. Feng, H. Y. Li, and L. L. Li, "An improved whale algorithm for support vector machine prediction of photovoltaic power generation," *Symmetry*, vol. 13, no. 2, pp. 1–26, 2021, doi: 10.3390/sym13020212.

[35] A. Guezzaz *et al.*, "A global intrusion detection system using PcapSockS sniffer and multilayer perceptron classifier," *International Journal of Network Security*, vol. 21, no. 3, pp. 438–450, 2019.

[36] Z. Car, S. B. Šegota, N. Anđelić, I. Lorencin, and V. Mrzljak, "Modeling the spread of COVID-19 infection using a multilayer perceptron," *Computational and Mathematical Methods in Medicine*, vol. 2020, 2020, doi: 10.1155/2020/5714714.

[37] P. Mohapatra *et al.*, "Artificial neural network based prediction and optimization of centelloside content in centella asiatica: A comparison between multilayer perceptron (MLP) and radial basis function (RBF) algorithms for soil and climatic parameter," *South African Journal of Botany*, vol. 160, pp. 571–585, 2023, doi: 10.1016/j.sajb.2023.07.019.

[41] S. Nosratabadi, S. F. Ardabili, Z. Lakner, C. Makó, and A. Mosavi, "Prediction of food production using machine learning algorithms of multilayer perceptron and ANFIS," *SSRN Electronic Journal*, 2021, doi: 10.2139/ssrn.3836565.

[38] B. J. Saleh, A. Y. F. Saedi, A. T. Q. A. -Aqbi, and L. Salman, "Analysis of weka data mining techniques for heart disease prediction system," *International Journal of Medical Reviews*, vol. 7, no. 1, 2020, doi: 10.30491/ijmr.2020.221474.1078.

[39] K. Preet, S. Attwal, and A. S. Dhiman, "Exploring data mining tool-weka and using weka to build and evaluate predictive models," *Advances and Applications in Mathematical Sciences*, vol. 19, no. 6, pp. 451–469, 2020.

[40] K. A. Shakil, S. Anis, and M. Alam, "Dengue disease prediction using weka data mining tool," *arxiv-Computer Science*, Feb. 2015, pp. 1-26, doi: 10.48550/arXiv.1502.05167

[42] B. J. Erickson and F. Kitamura, "Magician's corner: 9. performance metrics for machine learning models," *Radiology: Artificial Intelligence*, vol. 3, no. 3, May 2021, doi: 10.1148/ryai.2021200126.

[43] O. Almomani, A. Alsaaidah, A. A. A. Shareha, A. Alzaqebah, and M. Almomani, "Performance evaluation of machine learning classifiers for predicting denial-of-service attack in internet of things," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 1, 2024, doi: 10.14569/IJACSA.2024.0150125.

[44] G. M. Foody, "Sample size determination for image classification accuracy assessment and comparison," *International Journal of Remote Sensing*, vol. 30, no. 20, pp. 5273–5291, Sep. 2009, doi: 10.1080/01431160903130937.

[45] D. Krstinić, M. Braović, L. Šerić, and D. B. -Štulić, "Multi-label classifier performance evaluation with confusion matrix," in *Computer Science & Information Technology*, AIRCC Publishing Corporation, Jun. 2020, pp. 01–14, doi: 10.5121/csit.2020.100801.

[46] N. J. R. Maria and R. Pankaja, "Performance analysis of text classification algorithms using confusion matrix," *International Journal of Engineering and Technical Research (IJETR)*, vol. 6, no. 4, pp. 75–78, 2016.

[47] F. Rahmad, Y. Suryanto, and K. Ramli, “Performance comparison of anti-spam technology using confusion matrix classification,” *IOP Conference Series: Materials Science and Engineering*, vol. 879, no. 1, 2020, doi: 10.1088/1757-899X/879/1/012076.

[48] H. Yun, “Prediction model of algal blooms using logistic regression and confusion matrix,” *International Journal of Electrical and Computer Engineering*, vol. 11, no. 3, pp. 2407–2413, 2021, doi: 10.11591/ijece.v11i3.pp2407-2413.

BIOGRAPHIES OF AUTHORS

Areen Arabiat obtained B.Sc. in Computer Engineering in 2005 from al Balqa Applied University, and her M.Sc. in Intelligent Transportation Systems from Al Ahliyya Amman University in 2022. She is currently a computer lab supervisor at the Faculty of Engineering/Al-Ahliyya Amman University since 2013. Her research interests are focused on the following areas: machine learning, data mining, artificial intelligence, and image processing. She can be contacted at email: a.arabiat@ammanu.edu.jo.

Mohammad Hassan has completed his Ph.D. from Baku State University, Azerbaijan. He is an Associate Professor in the Department of Computer Engineering at the Faculty of Engineering at Al-Ahliyya Amman University. He is a member of the Jordanian Engineering Association. He has published numerous research papers in various journals and conferences, covering topics such as machine learning, computer networks, intelligent transportation systems, and mobile learning adaptation models. He can be contacted at email: mhassan@ammanu.edu.jo.

Omar Almomani received his Bachelor's and master’s degrees in Telecommunication Technology from the Institute of Information Technology at the University of Sindh in 2002 and 2003 respectively. He received his Ph.D. from the University Utara Malaysia in computer network and security in 2010. Currently, he is a professor in the Department of Networks and Cybersecurity, Faculty of Information Technology at Al-Ahliyya Amman University. His research interests involve network performance, network quality of service (QoS), IoT, network modeling and simulation, network and cyber security, and machine learning. He can be contacted at email: at almomani81@yahoo.com.